Briefings in Bioinformatics, 17(2), 2016, 193–203

doi: 10.1093/bib/bbv033
Advance Access Publication Date: 9 June 2015
Paper

Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks

Xiangxiang Zeng, Xuan Zhang and Quan Zou

Corresponding author: Quan Zou, 422# Simingnan Road, Department of Computer Science, Xiamen Univeristy, Xiamen, 361005. P.R.China. Tel.: +86-592-2580333; Fax: +86-5922580033; E-mail: zouquan@xmu.edu.cn

Abstract

MicroRNAs (miRNA) play critical roles in regulating gene expressions at the posttranscriptional levels. The prediction of disease-related miRNA is vital to the further investigation of miRNA's involvement in the pathogenesis of disease. In previous years, biological experimentation is the main method used to identify whether miRNA was associated with a given disease. With increasing biological information and the appearance of new miRNAs every year, experimental identification of disease-related miRNAs poses considerable difficulties (e.g. time-consumption and high cost). Because of the limitations of experimental methods in determining the relationship between miRNAs and diseases, computational methods have been proposed. A key to predict potential disease-related miRNA based on networks is the calculation of similarity among diseases and miRNA over the networks. Different strategies lead to different results. In this review, we summarize the existing computational approaches and present the confronted difficulties that help understand the research status. We also discuss the principles, efficiency and differences among these methods. The comprehensive comparison and discussion elucidated in this work provide constructive insights into the matter.

Key words: disease miRNA prediction; network similarity; biological database

Introduction

MicroRNAs (miRNAs) are a class of small, endogenous, single-stranded, noncoding RNAs (∼22 nucleotides) that mainly repress the expression of target mRNAs at the posttranscriptional level by binding to the 3′-untranslated region of target mRNAs through sequence-specific base pairing, resulting in target mRNAs cleavage or translation inhibition [1–3]. Increasing evidence suggests that miRNAs may function as positive regulators at the posttranscriptional level, which is critical for development of diseases [4, 5]. The lin-4 and let-7 in *Caenorhabditis elegans* were the first miRNAs to be discovered [6]. Since then, using various experimental methods, numerous miRNAs have been identified. As of June 2014, investigators have discovered and documented 28 645 miRNA entries, which can be run through the miRBase. Increasing evidence suggests that the mutation of miRNA, the dysfunction of miRNA biogenesis and the dysregulation of miRNAs and miRNA target genes may lead to various diseases. As such, miRNA, first discovered in C. *elegans* lin-4, encodes a 22-nt RNA fragment length and regulates the expression of its target genes lin-14 [7] and lin-28 [8] and plays an important regulatory role in the development of nematode larvae [6]. The second discovered miRNA let-7 regulates its target gene lin-41 [9] and hbl-1 [10].

Herein, identifying the interactions between miRNAs and diseases is a crucial problem. Applying only experimental methods to discover such a correlation poses many bottlenecks such as lengthy experimental periods, high equipment requirements and high cost. With the emergence of numerous miRNAs, various databases have been presented to store meaningful information about these RNA molecules. Based on existing data, as a supplement of experimental methods, computational approaches were

Xiangxiang Zeng is an assistant professor in Xiamen University. He is a member of IEEE and ACM. His research interests include systems biology, heterogeneous networks and link prediction.
Xuan Zhang is a graduate student in Xiamen University. Her research interest is disease microRNA prediction.
Quan Zou is an associate professor in Xiamen University. He is a member of IEEE and ACM. His research interests include bioinformatics and data mining.

Submitted: 10 March 2015; **Received (in revised form):** 21 April 2015

proposed to solve the problem. Computational means obtain the potential associations between miRNAs and diseases in a short time and greatly reduce the experimental workload. However, many challenges also confront these novel approaches. First, we have only positive samples without negative samples. Because, through several experimental means, we can prove an miRNA is related with a disease, but cannot prove an miRNA is absolutely unrelated with a disease. This results in fewer positive associations and a vast amount of unlabeled samples. Second, when a new miRNA is discovered, there is no information about it so that the computational approaches can hardly predict the relationship of this RNA molecule with any diseases.

The rest of this article is organized into the following four sections: Section 2 summarizes the databases used to predict certain miRNA–disease associations; these databases can be downloaded from the specified links. Section 3 reviews previous approaches and divides them into two categories. The first category is based on the similarity network method; the other is based on the machine learning method. Finally, Section 4 concludes the review and discusses the advantages and disadvantages of previous approaches.

Database

In this section, we present related databases to help predict certain miRNA–disease associations.

miRNA–disease network

HMDD v2.0 and miR2Disease are databases that have long been developed and made publicly available. These information banks can be accessed through a web interface at http://cmbi.bjmu.edu.cn/hmdd and http://www.miR2Disease.org/, respectively. HMDD v2.0 is a database to experimentally support human miRNA and disease associations [11]. MiR2Disease is a manually curated database for miRNA deregulation in human disease [12]. Another special database, miRCancer, focuses on the miRNA–cancer relationship. ThemiRCancer is an miRNA–cancer association database constructed by text mining of the literature [13]. The dbDEMC is a database of differentially expressed miRNAs in human cancers [14].

Disease phenotype network

To obtain more comprehensive and complete networks, some researchers use the disease phenotype network, which is based on more information and easily attains good performance. Accordingly, certain scholars, via the gene–phenotype network, obtain a disease similar network. A more recent database for disease and gene association, named DGA, can be downloaded from http://dga.nubic.northwestern.edu. [15]. Genes associated with similar diseases show both high likelihood of physical interactions and higher expression profiling similarity for their transcripts, supporting the existence of distinct disease-specific functional modules [16]. Theoretically, this guarantees the feasibility of the aforementioned strategy. More generally, by text mining, the MimMiner database is widely used to classify >5000 human phenotypes contained in the Online Mendelian Inheritance in Man database (OMIM). Researchers found that the similarity between phenotypes reflects biological modules of interacting functionally related genes [17]. The phenotype association data can be downloaded at http://www.cmbi.ru.nl/MimMiner/. In addition, another database, MeSH, which is available at http://www.ncbi.nlm.nih.gov/, can be translated into disease sematic similarity information through some process to be described later [18].

miRNA association networks

The miRNA–miRNA relationship cannot be directly identified. Therefore, establishing a functional association between two different miRNAs, generally via their target-genes network, is useful for attaining good performance. Numerous miRNA-target gene databases are adopted. We roughly divide these databases into two categories arbitrarily designated ‘Category A’ and ‘Category B’ according to the source of their relation.

Category A. Database miRNAMap 2.0, a representative class in this category, contains experimentally verified miRNAs and miRNA target genes in human, mouse, rat and other metazoan genomes [18]. The miRNAMap 2.0 is now available at http://miRNAMap.mbc.nctu.edu.tw/. TarBase [19] and miRNA.org [20], which also belong to this category, can be download from http://www.diana.pcbi.upenn.edu/tarbase/and http://www.miRNA.org/.

Category B. Another database, miRBase Sequences, is the primary online repository for miRNA sequence data and annotation, which is a comprehensive new database of predicted miRNA target genes [21]. MiRBase is available at http://miRNA.sanger.ac.uk/. Similar databases in this category, such as PITA [22] miRGator [23], miRGen [24], PicTar [25], TargetScan [26], DIANA-microT [27], RNAhybrid [28], RNA22 [29] are often used to obtain the relationship between miRNA and its target genes. All these databases are available online. Based on prior work, the miRNA–miRNA functional similarity scores can be downloaded from http://cmbi.bjmu.edu.cn/misim/ [18]. Rfam [30], an RNA family database, may also be used when miRNA familiar information is considered.

Gene interaction network (OMIM)

Genes play an extremely important role in biological inheritance. To a certain extent, genes with similar functions are often associated with similar diseases. For this reason, a large amount of gene information is used to predict diseases. Gene–phenotype relationships, which can be obtained from the OMIM, are widely used [31]. Because OMIM does not routinely collect findings of lower significance or negative findings and because of the increase in the development of genetic association databases, GAD was built [32]. The related data are available at http://geneticassociationdb.nih.gov. Also, the association between diseases and genes can be obtained from the disease-metabolic subpathway network (DMSPN) [33].

Protein interaction network (PPI)

The protein interaction network can be indirectly used to resolve this issue. The protein–protein interaction (PPI) data were derived from the Human Protein Reference Database, which created proteomic information pertaining to human proteins [34]. To achieve satisfactory performance, some researchers also use such a network. Protein–disease associations can be obtained from the DISEASES database at http://diseases.jensenlab.org [35].

Computational methods to predict disease-related miRNA

To the best of our knowledge, this review is the first to review the state-of-the-art methods. Generally speaking, there are two main types of methods to address disease–miRNA prediction. One is experimental identification, which encounters many bottlenecks (e.g. it is time-consuming and expensive). Herein, computational means, proposed to supplement the experimental method, drastically reduce the number of candidate miRNAs. All these computational methods have one consistent purpose: predicting the relationship between disease and miRNAs. For ease of understanding, we state our issue intuitively in Figure 1. Some verified miRNA–disease associations are shown as straight lines, while potential relationships are shown as dotted lines. Disease and miRNA network data can be downloaded directly.

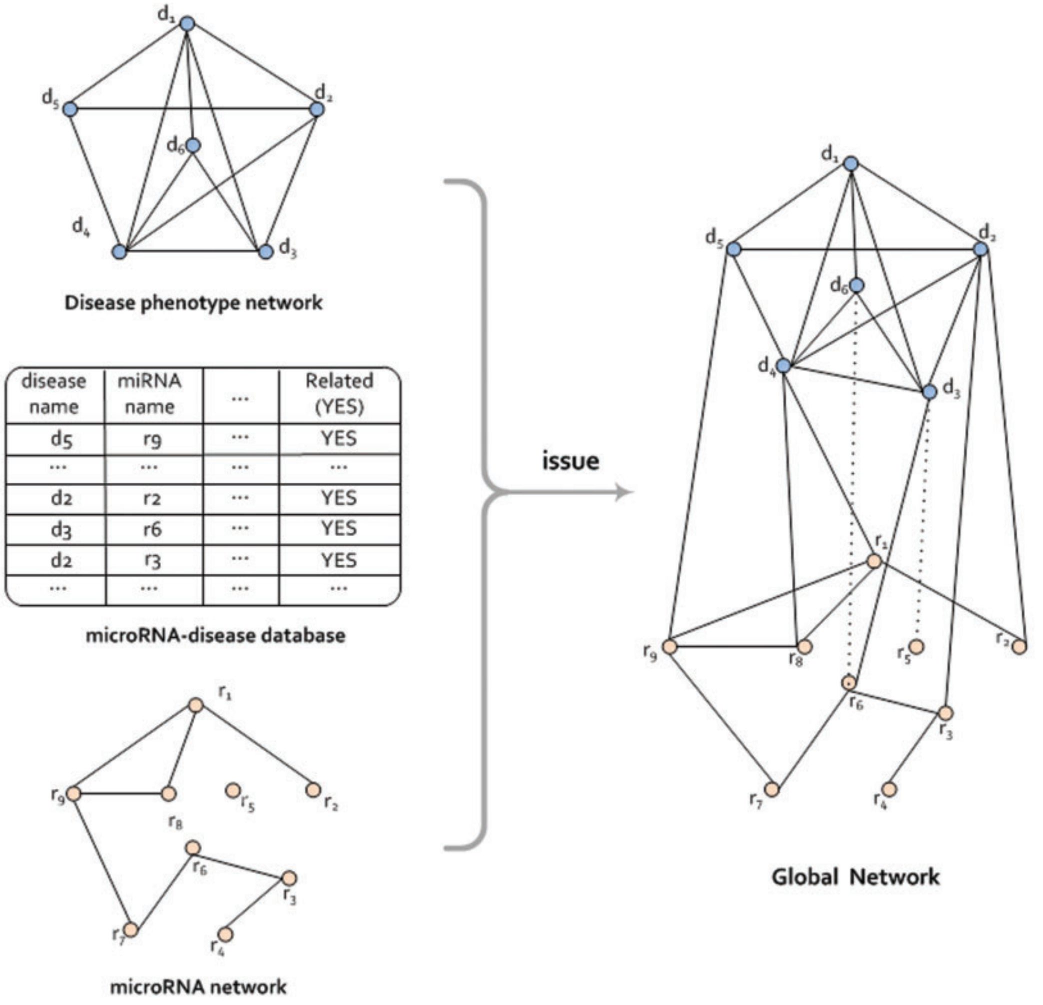
Our goal is to judge whether an miRNA is related to a specific disease. Of the massive methods, the key is how to define similarity scores (e.g. phenotype-miRNA similarity scores, miRNA similarity score, phenotype similarity score). Two biological assumptions are used to identify the associations between 两个生物学依据： miRNAs and diseases. Some researchers compute similarity score based on the tendency that 1 miRNAs are associated with phenotypically similar diseases. However, similarity score cannot be computed for diseases that have no known associated miRNAs. Whereas, methods based on the theory that 2 functionally related miRNAs likely tend to be associated with phenotypically similar diseases can work for all diseases [36]. Here, we introduce, in chronological order, the proposed approaches from the following two aspects: similarity measure approaches (see Section 3.1) and machine learning means (see Section 3.2).

Based on similarity measure methods

1. In 2009, Jiang *et al.* [36] proposed the first computational method, which focuses on how to prioritize disease miRNAs based on a human phenome-miRNAome network. In their method, Jiang, *et al.* first constructed a functionally related miRNA network and a human phenome-miRNAome network [36]. The human phenome-miRNAome network data can be downloaded from the database described above in Section 2. The miRNAome network was constructed according to the following two considerations: miRNA functional models information and shortest pathway between two miRNAs. Specifically, the threshold is set at 0.3 for the disease phenotype network data. If the similarity score of the two diseases is no less than 0.3, the diseases are considered to be related; when the similarity score is equal to 1, the two diseases are identical. After that, a scoring system is proposed to prioritize miRNA. Sorting all of the miRNAs based on this score, the top ranked miRNA has a higher probability to associate with a specific disease. Herein, for a disease of interest, *d*, the miRNA score is calculated by the cumulative hypergeometric distribution: 累积超几何分布？？

$$score = 1 - \sum_{i=m}^{M} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

N was the total number of miRNAs in entire miRNA similarity network data. *M* was the total number of miRNAs known to



disease name	miRNA name	...	Related (YES)
d5	r9	...	YES
...	...	...	...
d2	r2	...	YES
d3	r6	...	YES
d2	r3	...	YES
...	...	...	...

Figure 1. Illustration of the network by using a special example.

be related to disease *d*. *n* is the number of miRNAs in the corresponding miRNA module. *m* is the number of miRNAs associated with similar diseases and is found in the corresponding module. The scores, as calculated by the above formula, are shown in Figure 2.

For the method of Jiang *et al.* [37], the area under the receiver operating characteristic (ROC) curve (AUC) is 0.785, indicating that this method achieved reasonable performance in identifying known disease-related miRNAs. This performance can be improved by later methods.

2. In 2010, Jiang *et al.* [37] proposed a new method, based on genomic data integration, to solve the above problem, to some extent. The Naïve Bayes model was used to integrate multiple types of data resources and build a model to predict functional between genes [37]. The associations between disease and gene are denoted as vectors, V_d, while the associations between miRNA and target genes are denoted as vectors, V_m. For a given disease, Jiang *et al.* [37] computed the similarity score for every miRNA and ranked the scores from highest to lowest. There is a high probability that the top ranked miRNA score will predict a disease-related miRNA.

3. Chen *et al.* [38] applied random walk on the miRNA–miRNA functional similarity network and achieved an AUC of 86.17%. The random walk method, successfully applied in prioritizing candidate disease genes, randomly simulated a walker's transition from its current nodes to neighbors in the network starting at some given seed nodes [39]. There are three steps as follows: First, decide the initial probability of each mircoRNA; second, execute random walk on an miRNA functional similarity network; third, obtain the stable probability of random walk and rank the candidate miRNAs. For a given disease, d, Chen *et al.* 4.
[39] denoted that miRNAs are confirmed to be associated with the disease as seed nodes of random walk, and all the remaining miRNAs are candidates. They initialized (1) the probability value of nodes p(0) where the possibility value of every seed node is equal, and the sum is equal to 1 and (2) the possibility value of every candidate node whose value is equal to zero. A matrix, W, is available from the previous database, which consists of miRNA–miRNA functional similarity data. Here, in every iteration *k*, Chen *et al.* [39] allowed the restart of random walk at source nodes with probabilities *p(k)*, as a vector, representing the probability of the random walk at a node *i* at step *k*. Hence, the probability of random walk is defined as:

$$p(k+1) = (1-r)W_{P(k)} + rp(0)$$

The condition for terminating the iterations is that the number representing the change between p(k) and p(k + 1) is less than the cutoff value of 10-6. After completion of the iterations, this method obtains a stable probability, $p(\infty)$. Based on the value of $p(\infty)$, Chen *et al.* ranked the candidate miRNAs related to the disease of interest, d. The higher the score of an miRNA, the higher the probability that it is associated with a known disease, d. The RWRMDA algorithm, visualized in Figure 3, was tested by the leave-one-out cross-validation and had a satisfying performance compared with previous methods.

In 2013, Chen *et al.* [39] focused on computing a similarity and proposed a similarity-based method. Three strategies—MBSI (miRNA-based similarity inference), PBSI (phenotype-based similarity inference) and NetCBI (network-consistency-based inference)—based on different data network, achieved AUC values of 74.83, 54.02 and 80.66%, respectively [40]. To aid in the understanding of the above strategies, Chen *et al.* [40] introduced the following terminology: (1) two sets, M = {m1, m2, . . . ,mn} and P = {p1, p2, . . . ,pm}, to represent miRNAs and phenotypes; (2) a bipartite graph G (M, P, E), where E = {eij:mi∈M,pi∈P}. If an miRNA i has a relationship with a phenotype j, an edge eij is drawn in graph G. From the perspective of method implementation, we stored the graph, G, in an m × n adjacent matrix, A = {aij(i<n,j<m)}, where aij = 1 if an miRNA, i, and a phenotype, j, are verified as related; otherwise, aij = 0. (3) To store the data, three adjacency matrixes A(m × n), M(n × n) and P(m × m) can be downloaded from the above databases.

In the MBSI method, Chen *et al.* proposed that for a given disease, *d*, if an miRNA, *i*, is related to the disease, *d*, other miRNAs, which are similar with the miRNA, *i*, have high probabilities to regulate the disease, *d*. Based on this consideration, they computed a similarity score between m_i and p_i using the following *MBSIscore* formula:

$$MBSIscore_{ij} = \frac{\sum_{l=1,l\neq i}^{n} S(m_i, m_l)a_{lj}}{\sum_{l=1,l\neq i}^{n} S(m_i, m_l)}$$

Where $S(m_{i,}\ m_l) = M_{il}$, and M_{il} is an miRNA functional similarity value between m_i and m_l.

For the same consideration, they proposed that an miRNA, *i*, had a high probability to associate with the disease, which is similar to a disease, d, known to be linked to the miRNA, *i*. Herein, they used the *PBSIscore* formula to obtain a similarity score between m_i and p_i as follows:

$$PBSIscore_{ij} = \frac{\sum_{l=1,l\neq i}^{n} S(p_i, p_l)a_{lj}}{\sum_{l=1,l\neq i}^{n} S(p_i, p_l)}$$

where $S(p_i, p_l) = P_{il}$, and P_{il} is an miRNA functional similarity value between p_i and p_l.

Based on network consistency, to gain better performance, they integrated phenotype similarity network data and miRNA similarity network data. Based on the graph Laplacian of *M(n × n)*, they calculated the similarity scores of an miRNA *i* linked to other miRNAs. Normalizing *M* = M(:,i)/sum(M(:,I)), a vector *m* of the graph Laplacian is derived from the formula

$$\min_{\tilde{m}} \sum_{i,j} \overline{M}_{i,j}(\tilde{m}_i - \tilde{m}_j)^2 + \frac{1-a}{a}\sum_{i}(\tilde{m}_i - m_i)^2$$

To reduce unnecessary calculations, they found that a close solution to this formula is $\tilde{m} = (1-a)(I - a\overline{M})^{-1}$. Similarly, this graph Laplacian score is also applied to compute a phenotype vector, $\tilde{p} = (1-\beta)(I-\beta\overline{P})^{-1}$, where $\overline{P}$ is the normalization of *P* and $\alpha, \beta \in (0, 1)$ are parameters. A similarity score between an miRNA, *i*, with a phenotype, *j*, was computed via the Pearson correlation coefficient score by the following equation:

$$NetCBIscore(\tilde{m}, \tilde{p}, a) = corr(a\,\tilde{p}, \tilde{m})$$

where *a* is a vector from the matrix *A(m × n)*. A phenotype is more likely to link with an miRNA, whose NetCBIscore is the higher one.

5. In 2013, based on the RWR algorithm, an improved method was proposed by Shi *et al.* [40]. In identifying known cancer-related miRNAs for the nine human cancers, a complementary network, containing miRNA–targets networks, disease–genes networks and protein–protein networks achieved a satisfactory performance with an AUC ranging from 71.3 to 91.3% [40]. Based on previous biological knowledge that miRNAs regulated diseases through their target genes, the hypothesis was put forth that if an miRNA target gene is associated with a disease gene, there is a high probability that the miRNA will affect the disease. Specifically, Shi *et al.* mapped disease genes and miRNA target genes on the PPI network and obtained two ranked lists of genes, derived from the RWR algorithm with different seeds. As a part of their method, Shi *et al.* performed the following three-step process: Step 1: set the disease genes as seeds; Step 2: on the seeds in Step 1, do the gene set enrichment analysis to determine the ES score; Step 3: set the miRNA genes as seeds and repeat Step 2 for these seeds. The ES score (the sum of ES1 and ES2) is computed by the following equations and visualized in Figure 4:

$$ES_1 = \max\left(\sum_{g_j \in TG, j \le i} \sqrt{(N-n_1)/n_1} - \sum_{g_j \in TG, j \le i} \sqrt{n_1/(N-n_1)}\right)$$

$$ES_2 = \max\left(\sum_{g_j \in TG, j \le i} \sqrt{(N-n_2)/n_2} - \sum_{g_j \in TG, j \le i} \sqrt{n_2/(N-n_2)}\right)$$

$$ES = \beta ES_1 + (1-\beta) ES_2$$

Shi *et al.* [40] used the *P*-value to measure the significance of an association between an miRNA and a disease and set the threshold to determine whether an miRNA is related to a given disease, *d*, without ranking miRNAs.

6. Xuan *et al.* [41] proposed another novel method, HDMP, based on the weighted k most similar neighbors to predict disease-related miRNAs [41]. For an miRNA *u* and an miRNA *v*, the similarity of *Misim(u,v)* was computed. Xuan *et al.* denoted two disease sets, DT_u and DT_v, which contain all diseases related to miRNA *u* and *v*, respectively. Using the following equation, an *Misim(u,v)* value is calculated:

$$Misim(u,v) = \frac{\sum_{1 \le i \le |DT_u|} DS(d_i, DT_v) + \sum_{1 \le j \le |DT_u|} DS(d_j, DT_u)}{|DT_u| + |DT_v|}$$

where $|DT_u|$ is the number of diseases in the set DT_u and $S(d_i,DT_v)$ is the similarity score between a disease d_i and the diseases in set DT_v. (For greater detail, see the flow chart in Figure 5). To obtain the value of $S(d_i,DT_v)$, Xuan *et al.* denoted two kinds of similarity between the two diseases: (1) the semantic similarity score $SS(d_i, d_j)$ and (2) the phenotype similarity $PS(d_i, d_j)$. Specifically, these similarities are computed based on the fact that diseases from the same level have a different contribution value because of the different frequency of occurrence of diseases in the GAD. Incorporating the semantic similarity and the phenotype similarity, the similarity between d_i and d_j is defined as $DS(d_i, d_j)$. Herein, they set the similarity score between a disease d_1 and a set of diseases, DT_v, as the maximum similarity between the disease d_1 and a disease thought all diseases in the set DT_v, which can be expressed in the formula.

$$S(d_1, DT_v) = \max\left(SS(d_1, d_{j_1}), SS(d_1, d_{j_2}), \ldots, SS(d_1, d_{j_n})\right)$$

This method computed the similarity score between two diseases by the directed acyclic graph (DAG) showing the relationship of diseases in different level. Xuan *et al.* denoted a different contribution factor ($\Delta = 0.5$) with a different level. In the *0th* level, there is only the diseases itself, so they set the factor as 1. With the increasing of level, the contribution factor decreased exponentially. For the instance, the contribution factor in the *ith* is the Δ^i. An example is showed in Figure 6. The similarity score of two diseases is a ratio concerning the contribution value of same partition of two diseases and the contribution value of all associations about two diseases. Through the above description, the similarity scores of two miRNAs are computed and a symmetric functional similarity matrix is constructed. In addition, by considering the miRNA family and the cluster information, this method obtained a satisfactory performance.

7 In 2013, another protein-driven inference method was proposed by Mørk *et al.* [35], which predict miRNA–disease associations by using the linkage among miRNAs, proteins and diseases [35]. Specifically, for miRNA–disease associations, they downloaded miRNA-target data from the previous database and then mapped all miRNAs to miRBase identifiers and all targets to the ensemble protein identifiers using the STRING aliases file. Every association between an miRNA m and a protein p has a similarity score *T(m,p)*. For protein–disease associations, they downloaded the complete data set from the DISEASES database and obtain a similarity score *Z(d,p)* between a disease *d* and a protein *p*. The mapped data are available at http://mirpd.jensenlab.org. Two formulae are denoted to compute the similarity of miRNAs and diseases as follows. Note that this method is better than selecting miRNA–disease pairs randomly.

$$SCore_1 = \sum_{P^* \in P_M \cap P_D} T(m,p^*) Z(d,p^*)$$

$$SCore_2 = \max_{P^* \in P_M \cap P_D} T(m,p^*) Z(d,p^*)$$

8. Unlike the method has been introduced, Xu*et al.*'s [42] method focuses on the prioritization of cancer-related miRNAs by integrating phenotype associations of multiple diseases with matched miRNA and mRNA expression profiles, which obtained that the average AUC scores for 11 cancers were 75.84 and 73.59% based on different database miR2Disease and HMDD respectively [42]. Their method computed without using any prior information that whether miRNAs are related to a given disease. Based on the biological assumption, this method judged the relationship of an miRNA and a disease via computing the functional similarity score of miRNA-targets and disease genes from the three perspectives: the sub-ontology biological processes (BP) of the Gene Ontology (GO), the Kyoto Encyclopedia of Genes and Genomes (KEGG) and the topological structure of the protein–protein interaction network (PPIN) [12]. In the different consideration, they calculated an overlap score to store similarity scores between miRNA targets and disease genes for a specific disease, then they achieve the final functional similarity score by integrating three overlap scores. Based on the final score, potential miRNA associations were prioritized for the given cancer type.

Based on machine learning methods

1. In 2010, Xu *et al.* [43] applied the machine learning method to predict the relationship between miRNAs and diseases. This approach aimed to distinguish positive miRNA–disease associations from large-scale negative miRNA–disease associations. The primary focus of this method was to extract features from

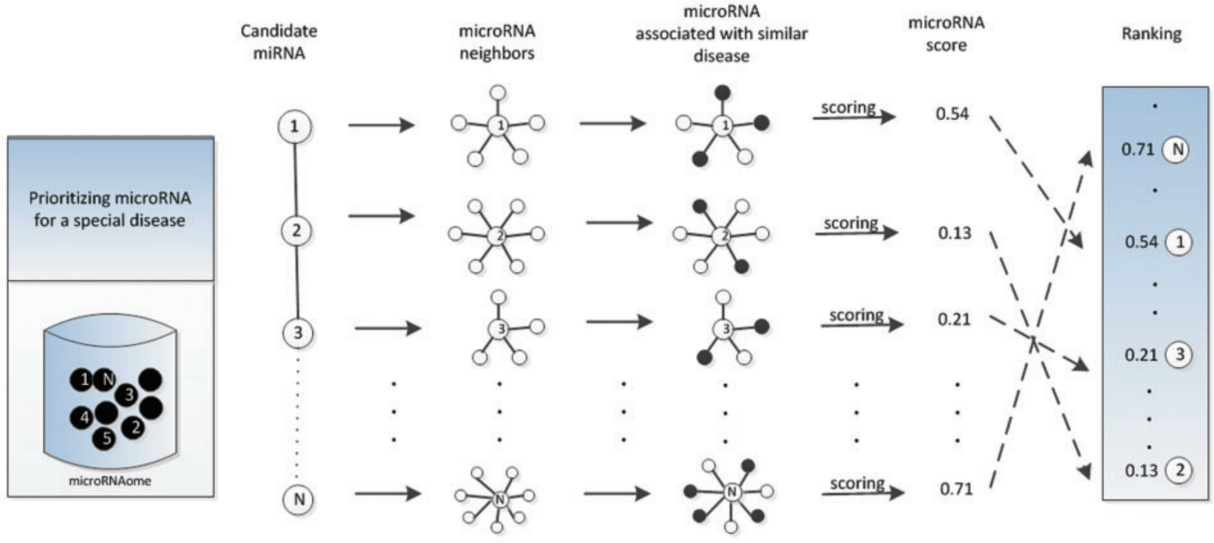


Figure 2. Flow path of algorithm.

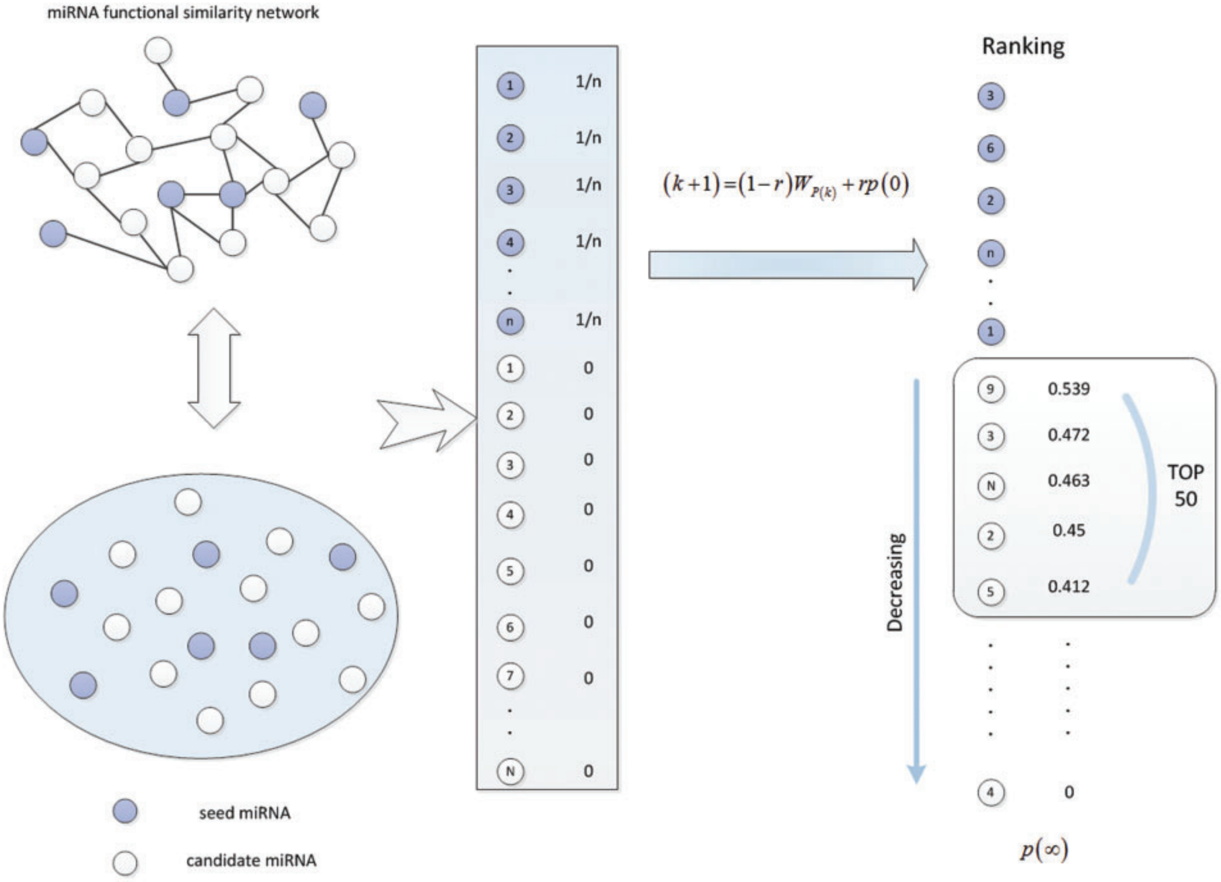


Figure 3. Flowchart of RWRMDA. A colour version of this figure is available at BIB online: http://bib.oxfordjournals.org.

the miRNA–disease network data and train a support vector machine (SVM) classifier, which obtained the AUC of up to 2. 0.8872 in cross-validation tests [43]. Jiang *et al.* [44] extracted different features to obtain an AUC value of 0.8884 in the 10-fold cross-validation [44]. Figure 7 shows the important flowchart of machine learning approaches. First, they constructed two feature vectors consisting of two feature sets. One feature set is primarily related to miRNA information, whereas the other is disease phenotype information.

Feature set primarily related to miRNA information. For the first feature set, each element value was calculated via the functional similarity score between an miRNA and the other miRNAs related to a given disease. To compute the functional similarity score between the two miRNAs, the miRNA

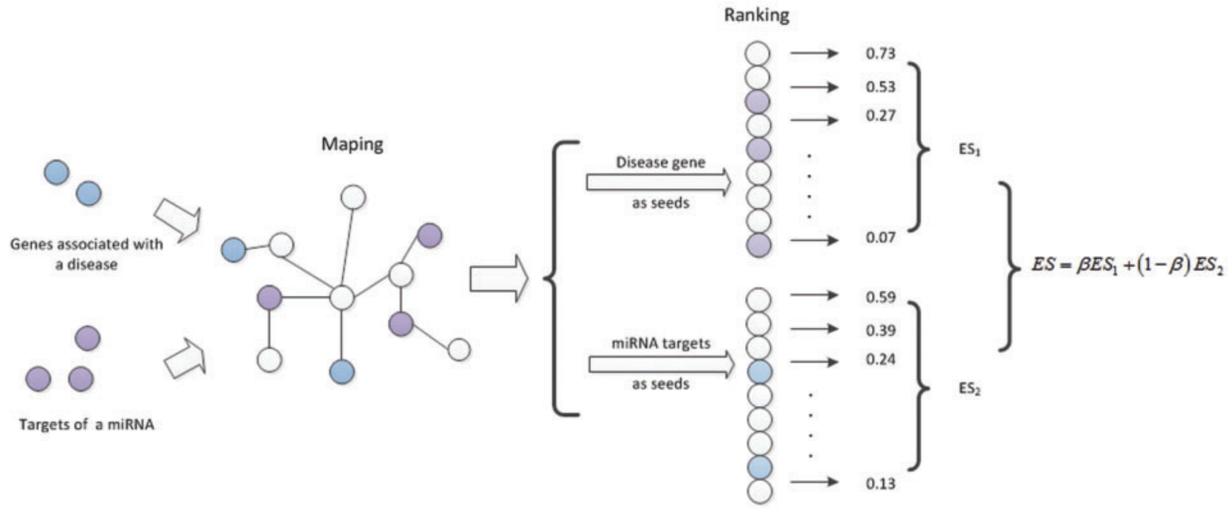


Figure 4. Calculating similarity score via mapping PPI network. A colour version of this figure is available at BIB online: http://bib.oxfordjournals.org.

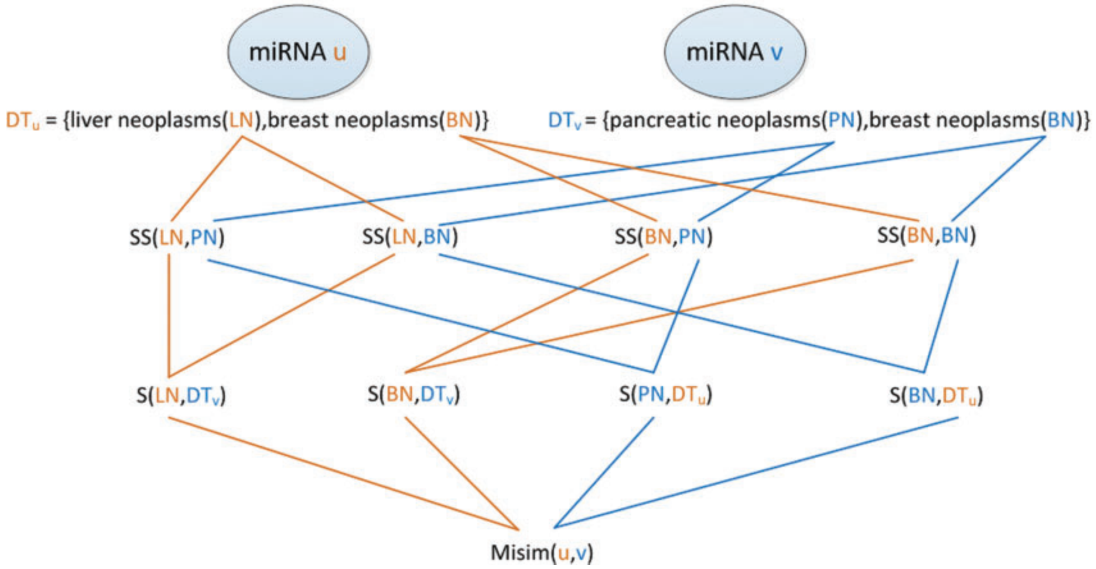

miRNA u
miRNA v
DT_u = {liver neoplasms(LN),breast neoplasms(BN)}
DT_v = {pancreatic neoplasms(PN),breast neoplasms(BN)}
SS(LN,PN)
SS(LN,BN)
SS(BN,PN)
SS(BN,BN)
S(LN,DT_v)
S(BN,DT_v)
S(PN,DT_u)
S(BN,DT_u)
Misim(u,v)

Figure 5. Calculating miRNA functional similarity. A colour version of this figure is available at BIB online: http://bib.oxfordjournals.org.

interaction data were downloaded from the aforementioned database and denoted as $\vec{V}_m = \{w_{m_i,g_1}, w_{m_i,g_2}, w_{m_i,g_3}, \ldots\}_{(i=1,2)}$, where $w_{mi,gj}$ is the interaction score predicted by the PITA [22] prediction algorithm. In this definition, the interaction scores of all genes in the human genome with miRNA *i* will be used. Note that the interaction score of the noninteracting miRNA-target was set to zero. Thereafter, they obtained an miRNA pair functional similarity score that is a Pearson's correlation coefficient between $\vec{V}_{m_1}$ and $\vec{V}_{m_2}$. The schematic illustration is shown in the Figure 7.

Disease phenotype information. In the second feature set, each element value was a disease phenotype similarity score, which was available from MimMiner, which was introduced in the database partition. If the score in the MimMiner was less than the noise threshold of 0.3, then we set the similarity score of the two diseases as zero. After the aforementioned operations, the feature vector of each miRNA–disease association contained 171(118 + 53) dimensions [38]. An n-fold cross-validation procedure was then performed. The positive and negative samples were randomly distributed into *n* parts. The $n-1$ parts samples were selected as the training set, and the one part sample was the testing part; the selection was repeated 10 times to ensure that each of the *n* parts was considered the testing part. Finally, the authors used the trained SVM classifier to predict the potential miRNA–disease relationships and classify the unlabeled data.

In 2014, considering the limitations in the previous methods, Regularized Least Squares for miRNA-Disease Association proposed by Chen *et al.* [45] was designed to construct a continuous classification function, which reflect the probability that each miRNA is related to a given disease; this function works for diseases without known related miRNAs [45]. This approach is a semi-supervised and global method, which simultaneously prioritizes associations for all the diseases without the negative samples. As successfully confirmed by experiments. This approach also correctly predicted the potential miRNA among the top three of 32 diseases and 34 disease–miRNA associations, as successfully confirmed by experiments. In this method, the

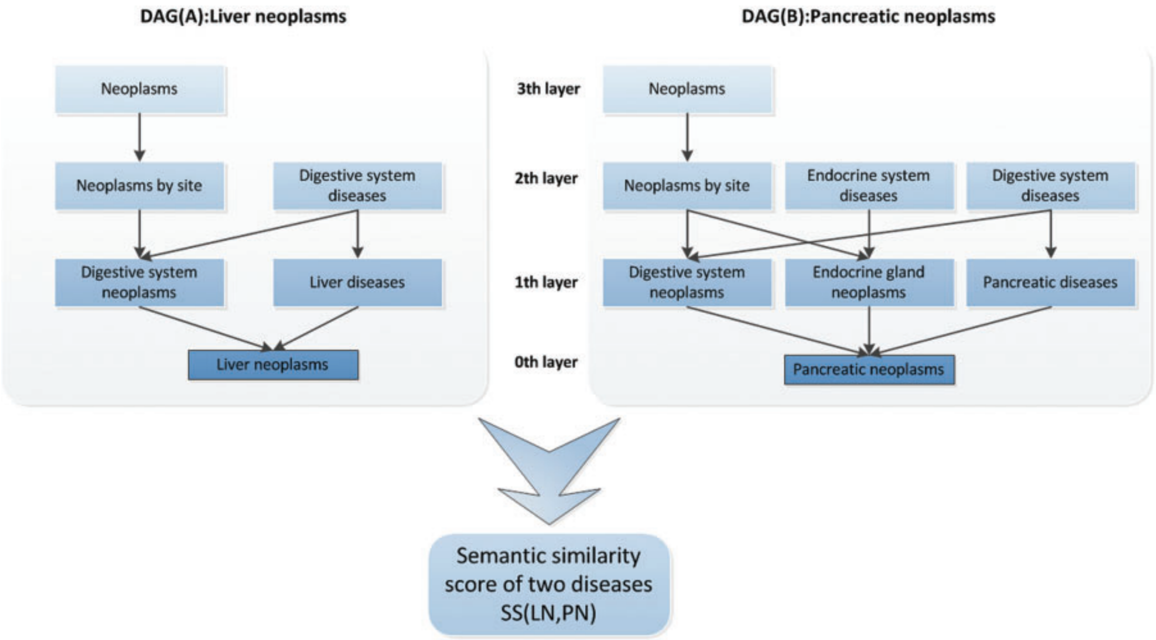

DAG(A):Liver neoplasms
DAG(B):Pancreatic neoplasms
3th layer
2th layer
1th layer
0th layer
Neoplasms
Neoplasms by site
Digestive system diseases
Digestive system neoplasms
Liver diseases
Liver neoplasms
Neoplasms
Neoplasms by site
Endocrine system diseases
Digestive system diseases
Digestive system neoplasms
Endocrine gland neoplasms
Pancreatic diseases
Pancreatic neoplasms
Semantic similarity score of two diseases SS(LN,PN)

Figure 6. Calculating disease's similarity via DAG. A colour version of this figure is available at BIB online: http://bib.oxfordjournals.org.

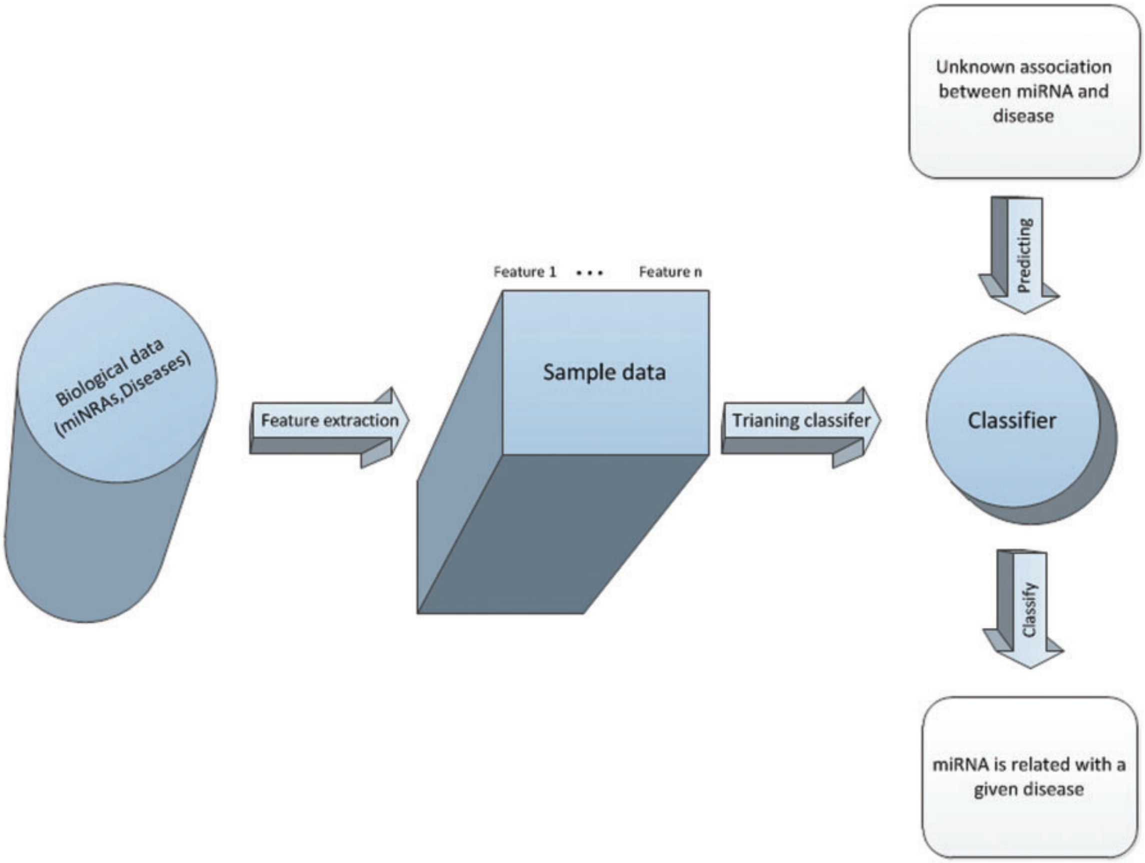

Unknown association between miRNA and disease
Predicting
Feature 1
• • •
Feature n
Biological data (miNRAs,Diseases)
Feature extraction
Sample data
Trianing classifer
Classifier
Classify
miRNA is related with a given disease

Figure 7. Flowchart of machine learning methods. A colour version of this figure is available at BIB online: http://bib.oxfordjournals.org.

researchers used the disease data through a technique similar to the HDMP method. To solve the problem of no negative samples, a semisupervised classifier was constructed under the framework of Regularized Least Squares, which was obtained by defining and minimizing a cost function. This cost function would be trained in the miRNA and disease networks, and the forward strategy was then used to optimize the classification function. Finally, the researchers obtained two different optimal classifiers in both miRNA and disease spaces. Both spaces simultaneously contributed to solve this problem, and a parameter

Table 1. Comparison of disease-related miRNA prediction methods

Method	Type of networks	Type of databases	Amount of data	AUC values	Type of test	Time	References
Jiang *et al.*	M-D	HMDD, miR2Diseas	270 (M-D assos)	75.80%	Cross-fold-validation	2009.8	[37]
	M-M	PITA, TargetScan, miRBase	145 872; 205 587 (M-G assos)				
	D-D	MimMiner					
Jiang *et al.*	M-D	miR2Diseas	62 miRNAS	–	Ranking	2010	[38]
	M-M	PITA	145 872 (M-G assos)				
	P-P	PPI	3 106 928 (P-P assos)				
Chen X *et al.*	M-D	HMDD, miR2Diseas, dbDEMC	1616 (M-D assos)	86.17%	Leave-one-out cross validation	2012.8	[39]
	M-M	misim					
Chen H *et al.*	M-D	HMDD, miR2Diseas	242 (M:99, D:51 assos)	80.66% (NetCBI)	Leave-one-out cross validation	2013.6	[40]
	M-M	misim		74.83% (MBSI)			
	D-D	MimMiner		54.2% (PBSI)			
Shi *et al.*	D-G	DMSPN	15 149 (D:412, G:2831)	71.3%–91.3 (In nine human cancers)	Cross-validation	2013	[41]
	M-G	miRanda, PicTar, TargetScan, DIANA-microT, RNA22, RNAhybrid, miRBase Targets	52 828 (M:566, g:8085)				
	PPI	PPI	36 867 (G-G assos, G:9453)				
Xuan *et al.*	M-D	HMDD	4739 (M:474, D:268)	82.47% (average in 18 diseases)	5 Fold-cross-validation	2013.8	[42]
	D-D	MimMiner	(D:5080)				
Mørk *et al.*	M-G	miRanda,TargetScan	14 599 (M:1169, P:1570, D:738)	–	Ranking	2013.11	[35]
	D-P	DISEASES					
Xu C *et al.*	M-D	HMDD, miR2Diseas	52 828 (M:566, G:8085)	75.84% (average)	Cross-validation	2014.7	[43]
	M-G	miRanda, PicTar, TargetScan, DIANA-microT, RNA22, RNAhybrid, miRBase Targets					
	D-D	MimMiner					
Xu J *et al.*	EP	miRanda, PicTar, TargetScan, DIANA-microT	484 043 (M:320, G:12713)	88.72%	Cross-validation	2011.7	[44]
	M-G						
Jiang *et al.*	M-D	HMDD, miR2Diseas	270 (D:53, M:118)	88.84%	10 fold-cross-validation	2013	[45]
	M-G	PITA, TargetScan, miRBase	145 872; 205 587 (M-G assos)				
	D-D	MimMiner					
Chen X *et al.*	M-D	HMDD	1616 (M-D assos)	84.50%	Leave-one-out cross validation	2014.6	[46]
	M-M	misim					
	D-D	DAG					

M-D, miRNA–disease associations; M-M, miRNA–miRNA associations; D-D, disease phenotype–disease phenotype associations; D-G, disease–gene associations; M-G, miRNA–target gene associations; D-P, disease–protein associations; M-P, miRNA–protein associations; P-P, protein–protein associations; EP, expression profiles.

was set to balance the final scores. A high score leads to a high probability of association between an miRNA and a given disease.

Comparisons with different approaches

In this section, we review the different computational methods to predict the disease–miRNA associations.

Table 1 provides an extensive overview of the methods that have been developed to date. In this table, the type of networks, databases used in the approaches and other fundamental information are compared. To understand the research progress. The same categories are discussed in a roughly chronological order to understand the research progress.

Table 1 presents an overview of the 11 approaches, which are continuously being developed and improved. This table also indicates that the prediction performance tested with ROC curve increased from 75.80 to 84.50%.

By applying the previous methods, the authors simply used the local network and limited information. Integrating more biological information networks guide researchers in considering the complete data and obtaining a satisfactory prediction performance. The method of Jiang *et al.* [36] can be improved based on several aspects. First, we consider that the miRNAs interaction network constructed using this method merely considered two aspects. More related information to build the miRNA network results in a more reliable miRNA association that predicts the disease-related miRNAs. Second, the Boolean network was used to store the phenotype–phenotype associations. Changing the Boolean network into a weight network is an improvement that can be implemented, and this change can lead to more detailed information. By replacing the Boolean network with weight network, Jiang *et al.* [37] improved the prediction performance of their study. Köhler *et al.* [46], Shi *et al.* [40] and Xuan *et al.* [41] obtained a high AUC value, but the performance of their methods depended on the selected parameters, including the restart probability, *P*-value and number of neighbors (k). Xuan *et al.* [41] used the similarity of the disease sets to compute the similarity score of miRNA. Obtaining a specific miRNA similarity score is significant and results in outstanding prediction performance. Several approaches, which have been applied by Köhler *et al.* [46] and Xuan *et al.* [41], could not predict diseases without the specified miRNAs. This problem was solved by using all the machine learning methods [43–45] and any of the non-machine learning methods of Jiang *et al.* [44], Chen *et al.* [39] and Xu *et al.* [43]. However, a few researchers may regard the unlabeled samples as negative in the application of machine learning methods. This procedure may lead to a low prediction performance. Chen *et al.* [45] recently addressed this issue and obtained better performance compared with the Random Walk with Restart for MiRNA–Disease Association techniques [38] and high-dose methylprednisolone [41].

Conclusion

Previous biological theories indicate that researchers can use computational methods to supplement biological experimental methods to predict the potential associations between miRNAs and diseases. Considerable biological interaction information is represented as networks, such as miRNA–disease, disease phenotype, miRNA association, gene interaction and protein interaction networks. The data of these networks can be downloaded directly. Integrating the network data to identify the associations between miRNAs and diseases is ideal and help researchers obtain better performance. Moreover, society can completely understand a biological system. One key issue in integrating a network is computing the similarities of the two nodes, particularly between an miRNA and a disease. In this review, we enumerate all types of approaches to address the similarity measure, and elucidate their respective advantages and disadvantages. In general, the similarity score can be determined by considering the neighbors, shortest pathway and familial information. Machine learning methods do not directly measure similarity score, but they consider such information through feature extraction.

In this review, we introduced several approaches to address the aforementioned issue and a few emerging problems that impeded the performance improvement of predictions. Future studies can focus on the following three aspects to improve the effects of forecasting performance. First, the miRNA interaction networks are not reliable because they were constructed using miRNA-target prediction tools, which contain numerous incorrect data. Hence, future scholars may develop more comprehensive data networks by using substantially related biological information. Second, future scholars may use updated data to obtain better prediction performance. Finally, future researchers must consider more biological knowledge to predict the association between miRNAs and diseases. Comprehensive and important information lead to accurate predictions.

Key Points

- Five existing databases are introduced to solve the problems in predicting disease-related miRNAs.
- Computational methods are proposed to address the issue, thereby complementing the experimental methods.
- We summarize these existing methods that have successfully predicted the association between miRNAs and diseases.
- Comparing and discussing the different aspects of these methods further investigate the process of predicting the interactions between miRNAs and diseases.

Funding

This work was supported by the Natural Science Foundation of China [grant numbers 61370010 and 61371179]; the Natural Science Foundation of Fujian Province of China [grant number 2014J01253]; and the Open Funding of Shanghai Key Laboratory of Intelligent Information Processing, China [grant number IIPL-2014-004].

References

1. Ambros V. The functions of animal microRNAs. *Nature* 2004; **431**(7006):350–5.
2. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 2004;**116**(2):281–97.
3. Meister G, Tuschl T. Mechanisms of gene silencing by double-stranded RNA. *Nature* 2004;**431**(7006):343–9.
4. Jopling CL, Yi M, Lancaster AM *et al.* Modulation of hepatitis C virus RNA abundance by a liver-specific MicroRNA. *Science* 2005;**309**(5740):1577–81.
5. Vasudevan S, Tong Y, Steitz JA. Switching from repression to activation: microRNAs can up-regulate translation. *Science* 2007;**318**(5858):1931–4.
6. Reinhart BJ, Slack FJ, Basson M *et al.* The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature* 2000;**403**(6772):901–6.
7. Wightman B, Ha I, Ruvkun G. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell* 1993;**75**(5): 855–62.
8. Moss EG, Lee RC, Ambros V. The Cold Shock Domain Protein LIN-28 Controls Developmental Timing in C. elegans and Is Regulated by the lin-4 RNA. *Cell* 1997;**88**(5):637–46.
9. Slack FJ, Basson M, Liu Z, Ambros V, Horvitz HR, Ruvkun G. The lin-41 RBCC Gene Acts in the C. elegans Heterochronic Pathway between the let-7 Regulatory RNA and the LIN-29 Transcription Factor. *Mol Cell* 2000;**5**(4):659–69.
10. Abrahante JE, Daul AL, Li M, Volk ML, Tennessen JM, Miller EA, Rougvie AE. The Caenorhabditis elegans hunchback-like Gene lin-57/hbl-1 Controls Developmental Time and Is Regulated by MicroRNAs. *Dev Cell* 2003;**4**(5):625–37.
11. Li Y, Qiu C, Tu J *et al.* HMDD v2. 0: a database for experimentally supported human microRNA and disease associations. *Nucleic Acids Res* 2013;**D42**:D1070–4.
12. Jiang Q, Wang Y, Hao Y *et al.* miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res* 2009;**37**(Suppl 1):D98–104.
13. Xie B, Ding Q, Han H *et al.* miRCancer: a microRNA–cancer association database constructed by text mining on literature. *Bioinformatics* 2013;**29**:638–44.
14. Yang Z, Ren F, Liu C *et al.* dbDEMC: a database of differentially expressed miRNAs in human cancers. *BMC Genomics* 2010; **11**(Suppl 4):S5.
15. Peng K, Xu W, Zheng J *et al.* The disease and gene annotations (DGA): an annotation resource for human disease. *Nucleic Acids Res* 2013;**41**:D553–60.

16. Goh K-I, Cusick ME, Valle D *et al*. The human disease network. *Proc Natl Acad Sci USA* 2007;**104**(21):8685–90.
17. Van Driel MA, Bruggeman J, Vriend G *et al*. A text-mining analysis of the human phenome. *Eur J Hum Genet* 2006;**14**(5):535–42.
18. Wang D, Wang J, Lu M *et al*. Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics* 2010;**26**(13): 1644–50.
19. Sethupathy P, Corda B, Hatzigeorgiou AG. TarBase: A comprehensive database of experimentally supported animal microRNA targets. *RNA* 2006;**12**(2):192–7.
20. Betel D, Wilson M, Gabow A *et al*. The microRNA. org resource: targets and expression. *Nucleic Acids Res* 2008;**36**(Suppl 1): D149–53.
21. Griffiths-Jones S, Grocock RJ, Van Dongen S *et al*. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res* 2006;**34**(Suppl 1):D140–4.
22. Kertesz M, Iovino N, Unnerstall U *et al*. The role of site accessibility in microRNA target recognition. *Nat Genet* 2007;**39**(10): 1278–84.
23. Nam S, Kim B, Shin S *et al*. miRGator: an integrated system for functional annotation of microRNAs. *Nucleic Acids Res* 2008; **36**(Suppl 1):D159–64.
24. Megraw M, Sethupathy P, Corda B *et al*. miRGen: a database for the study of animal microRNA genomic organization and function. *Nucleic Acids Res* 2007;**35**(Suppl 1):D149–55.
25. Krek A, Grün D, Poy MN *et al*. Combinatorial microRNA target predictions. *Nat Genet* 2005;**37**(5):495–500.
26. Grimson A, Farh KK-H, Johnston WK *et al*. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell* 2007;**27**(1):91–105.
27. Maragkakis M, Reczko M, Simossis VA *et al*. DIANA-microT web server: elucidating microRNA functions through target prediction. *Nucleic Acids Res* 2009;**37**:W273–6.
28. Krüger J, Rehmsmeier M. RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Res* 2006; **34**(Suppl 2):W451–4.
29. Miranda KC, Huynh T, Tay Y *et al*. A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell* 2006;**126**(6):1203–17.
30. Griffiths-Jones S, Bateman A, Marshall M *et al*. Rfam: an RNA family database. *Nucleic Acids Res* 2003;**31**(1):439–41.
31. Hamosh A, Scott AF, Amberger JS *et al*. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 2005;**33**(Suppl 1):D514–17.
32. Becker KG, Barnes KC, Bright TJ *et al*. The genetic association database. *Nat Genet* 2004;**36**(5):431–2.
33. Li X, Li C, Shang D *et al*. The implications of relationships between human diseases and metabolic subpathways. *PLoS One* 2011;**6**(6):e21131.
34. Prasad TK, Goel R, Kandasamy K *et al*. Human protein reference database—2009 update. *Nucleic Acids Res* 2009;**37**(Suppl 1):D767–72.
35. Mørk S, Pletscher-Frankild S, Caro AP *et al*. Protein-driven inference of miRNA–disease associations. *Bioinformatics* 2014; **30**:392–7.
36. Jiang Q, Hao Y, Wang G *et al*. Prioritization of disease microRNAs through a human phenome-microRNAome network. *BMC Syst Biol* 2010;**4**(Suppl 1):S2.
37. Jiang Q, Wang G, Wang Y: An approach for prioritizing disease-related microRNAs based on genomic data integration. In: *Biomedical Engineering and Informatics (BMEI), 2010 3rd International Conference on: 2010*. IEEE, 2270–4.
38. Chen X, Liu M-X, Yan G-Y. RWRMDA: predicting novel human microRNA–disease associations. *Mol Biosyst* 2012;**8**(10):2792–8.
39. Chen H, Zhang Z. Similarity-based methods for potential human microRNA-disease association prediction. *BMC Med Genomics* 2013;**6**(1):12.
40. Shi H, Xu J, Zhang G *et al*. Walking the interactome to identify human miRNA-disease associations through the functional link between miRNA targets and disease genes. *BMC Syst Biol* 2013;**7**(1):101.
41. Xuan P, Han K, Guo M *et al*. Prediction of microRNAs associated with human diseases based on weighted k most similar neighbors. *PloS One* 2013;**8**(8):e70204.
42. Xu C, Ping Y, Li X *et al*. Prioritizing candidate disease miRNAs by integrating phenotype associations of multiple diseases with matched miRNA and mRNA expression profiles. *Mol Biosyst* 2014;**10**:2800–9.
43. Xu J, Li C-X, Lv J-Y *et al*. Prioritizing candidate disease miRNAs by topological features in the mirna target–dysregulated network: Case study of prostate cancer. *Mol Cancer Ther* 2011;**10**(10):1857–66.
44. Jiang Q, Wang G, Jin S *et al*. Predicting human microRNA–disease associations based on support vector machine. *Int J Data Mining Bioinform* 2013;**8**(3):282–93.
45. Chen X, Yan G-Y. Semi-supervised learning for potential human microRNA-disease associations inference. *Sci Rep* 2014;**4**:5501.
46. Köhler S, Bauer S, Horn D *et al*. Walking the interactome for prioritization of candidate disease genes. *Am J Hum Genet* 2008;**82**(4):949–58.